

## Convergent and distributed effects of the 3q29 deletion on the human neural transcriptome

### Supplemental Information

Table of contents:

<b>Extended Methods</b> .....	3-15
Overview.....	3
Weighted gene co-expression network analysis (WGCNA).....	4-6
<i>Samples used for network construction</i> .....	4
<i>Data cleaning</i> .....	4-5
<i>Network construction and module detection</i> .....	5-6
Module preservation and quality analysis.....	7-8
Functional characterization of network modules harboring 3q29 genes.....	8-10
Identification of meta-modules harboring 3q29 genes.....	10
Determination of module membership strength for 3q29 genes.....	10-11
Identification of prioritized driver genes.....	11-13
Functional characterization of prioritized driver genes.....	13
Proof of concept study for testing the validity of WGCNA-based predictions.....	13-15
<i>Overview</i> .....	13
<i>RNA-sequencing in mouse cortex</i> .....	13
<i>Differential gene expression analysis</i> .....	14
<i>Comparison of empirically identified DEGs and network-derived predictions</i> .....	14-15
<b>Extended Results</b> .....	15-26
Unbiased gene co-expression network analysis reveals convergent and distributed effects of 3q29 interval genes across the adult human cortical transcriptome.....	15-16
Pathway analysis points to functional involvement of the 3q29 locus in nervous-system functions and core aspects of cell biology.....	16-20
Network modules found to harbor 3q29 interval genes are robust and strongly reproducible in an independent test dataset.....	20-21
<i>UBXN7</i> is a highly connected cortical hub-gene predicted to play a crucial role in the neuropsychiatric sequela of 3q29Del.....	21-22
Nine 3q29 interval genes form transcriptomic subnetworks enriched for known SZ, ASD and IDD-risk genes.....	22-25
Disease-relevant driver genes prioritized by network analysis load onto key biological pathways linked to neuropsychiatric disorders.....	25-26
Network-derived targets predict differentially expressed genes in the mouse model of 3q29Del.....	26
<b>Availability of Data &amp; Materials</b> .....	27
<b>Supplemental References</b> .....	28-34
<b>Supplemental Figures</b> (separate PDF file)	
Fig. S1. Tissue sample attributes and donor phenotypes of the GTEx dataset used for reference network construction.	
Fig. S2. Pre-processing of the reference dataset: outlier removal.	
Fig. S3. Determination of the soft-thresholding power ( $\beta$ ) in WGCNA.	
Fig. S4. Determination of network reproducibility and module preservation in an independent test dataset.	

Fig. S5. Individual module-preservation and quality statistics underlying composite  $Z_{\text{summary}}$  scores.

Fig. S6. Publication numbers for 3q29 genes and historic schizophrenia spectrum disorder candidate genes.

Fig. S7. Overlap between known protein-protein interactions (PPI) and gene co-expression patterns of 3q29 interval genes.

Fig. S8. STRING protein-protein interaction (PPI) networks of genes co-expressed in 3q29 modules.

Fig. S9. Amalgamated illustration of top biological processes / pathways enriched in 3q29 modules.

### **Supplemental Tables** (separate multi-sheet Excel file)

#### Supplemental Tables 1 (separate multi-sheet Excel file)

Table S1.1. Sample attributes and donor phenotypes for test and reference networks.

Table S1.2. Network-derived gene-module pairs.

Table S1.3. Eigengene-based module connectivity measures (kME).

Table S1.4. Module preservation and quality statistics for network validation.

Table S1.5. Functional enrichment results for network modules harboring 3q29 interval genes.

Table S1.6. Cortical mRNA expression summaries of 3q29 interval genes obtained from the Human Protein Atlas.

Table S1.7. Known binary protein-protein interactions of 3q29 interval genes curated from the Human Reference Protein Interactome (HuRI) Database.

Table S1.8. Protein-protein associations in gene co-expression network modules harboring 3q29 interval genes.

#### Supplemental Tables 2 (separate multi-sheet Excel file)

Table S2.1. Network-derived top intramodular neighbors of 3q29 interval genes.

Table S2.2. Curated gene-sets for disease association analysis.

Table S2.3. Network-derived prioritized driver genes & hypergeometric test results. assessing the overlap between curated gene-sets and top neighbors of 3q29 interval genes.

Table S2.4. Functional characterization of prioritized driver genes.

Table S2.5. Overlap between network-derived targets and differentially expressed protein coding genes in the 3q29Del mouse model.

## **Extended Methods**

### **Overview:**

For this analysis, we selected the most appropriate publicly available transcriptomic dataset and cleaned and processed the data before constructing an unsupervised gene co-expression network that can group functionally correlated genes into modules [1]. Genes participating in the same molecular and biological pathways tend to show correlated expression (co-expression) with each other, as they are expressed under the control of a coordinated transcriptional regulatory system [2]. Constructed on this principle, the transcriptomic network identified in this study amounts to a systems-level representation of gene-gene relationships in the healthy adult human prefrontal cortex (PFC), providing novel insights into network-level operations of understudied genes located in the recurrent 3q29 deletion (3q29Del) locus.

Given the strong genetic link between 3q29Del and risk for schizophrenia spectrum disorders (SZ; estimated odds ratio >40) [3], we placed our focus on revealing the co-regulation patterns of 3q29 interval genes as a function of their expression similarity during adulthood. This is a period when SZ typically manifests itself diagnostically, with peak onset in late adolescence (age 18-20 years) and early adulthood (age 20 to 40 years) [4], and a substantial proportion of patients becoming ill during middle adulthood (age 40 to 60 years) [5]. We also focused our analysis on uncovering gene-gene relationships in the PFC: a brain region that subserves a diverse range of cognitive and emotional operations, is implicated in the etiology of SZ and may be particularly vulnerable to the effects of genetic disruption due to its protracted development [6]. To validate the reproducibility of this transcriptomic network, we used an independent test dataset that was demographically comparable to our reference dataset and assessed the preservation (i.e., how well-defined modules are in an independent test dataset) of each identified network module by permutation tests. Using a similar approach, we also assessed the quality (i.e., how well-defined modules are relative to the background) of each network module in repeated random splits of our reference dataset.

We next identified and interrogated the modules that were found to harbor at least one 3q29 interval gene for (1) functional enrichment, (2) higher-level organization into meta-modules, (3) highly connected “hub” genes and (4) “top” network neighbors of 3q29 interval genes. To generate testable hypotheses about which 3q29 interval genes may be causally linked to the major neuropsychiatric phenotypes of 3q29Del, we tested whether the top network neighbors of individual 3q29 interval genes show significant overrepresentation of known risk genes for SZ, or for two other clinical phenotypes associated with 3q29Del: autism spectrum disorders (ASD) and intellectual / developmental disability (IDD). ASD and IDD share genetic risk and pathogenic mechanisms with SZ, despite differences in the timing of clinical symptoms [7]. Based on the identified enrichment patterns, we developed a list of “prioritized driver genes”, consisting of select 3q29 interval genes and their tightly correlated SZ, ASD and IDD-associated top network neighbors, which were subsequently found to converge onto canonical biological pathways.

We predict that dysregulation of these prioritized molecular targets, including a hub gene within the 3q29 interval, and their associated biological pathways may be implicated in the neuropsychiatric phenotype in 3q29Del syndrome. Overall, our findings highlight the advantage to using unbiased systems approaches that integrate gene-level information into a higher-order, network-level framework to infer gene function, particularly for understudied genes with disease relevance. The molecular pathways and the prioritized gene-set identified in this study will inform new directions of neurobiological inquiry that can mechanistically connect 3q29Del with severe mental illness. Our approach also highlights the study of normal function in non-pathological post-mortem tissue to further our understanding of psychiatric genetics, especially for rare genetic syndromes like 3q29Del, where access to post-mortem neural tissue from carriers is unavailable or limited. The steps described in this overview are described in great detail below.

## Weighted gene co-expression network analysis (WGCNA)

### *Samples used for network construction:*

To interrogate the functional consequences of the 3q29Del at the individual gene level, we employed unbiased weighted gene co-expression network analysis (WGCNA package [1, 8] in R, version 1.68) and organized the adult human cortical transcriptome into clusters (modules) of highly correlated genes (nodes). The network was constructed on publicly available RNA-Seq data obtained from the Genotype Tissue Expression Project (GTEx) [9] (Supp. Fig. S1), the largest multi-tissue open-data initiative using postmortem samples from human donors. 113 non-pathological post-mortem samples from the PFC (Brodmann Area 9) of male and female adults (age range = 20-79) with no known history of psychiatric or neurological disorder were included in this study. Transcriptome profiling was performed using Illumina TruSeq RNA sequencing as described in GTEx Consortium et al. (2013, 2015) [9]. The GTEx data (release version 6) used for the analyses described in this manuscript were downloaded from the GTEx portal (<http://www.gtexportal.org/home/datasets/>) on 01/08/2019; the corresponding dbGaP study accession number is phs000424.v6.p1.

### *Data cleaning:*

Six outlier samples were removed from the original dataset prior to network construction to prevent a potential outlier-driven bias in network structure and module detection (Supp. Fig. S2). Inter-sample correlation (ISC) was used as an unbiased statistical diagnostic for identifying outliers with divergent gene expression profiles, defined as the Pearson's correlation between pairs of samples across the expression levels of all detected genes. Samples with a mean ISC greater than two standard deviations away from the mean of the sample-set were flagged as outliers and removed (as described in Oldham et al., 2008 [10]), reducing the sample size to 107. Details on sample attributes and donor phenotypes are available in Table S1.1 and Supp. Fig. S1.

Protein-coding genes were extracted from the dataset based on GENCODE v19 annotations for gene-type [11], followed by  $\log_2$  transformation of gene expression values. For genes indexed by multiple splice variants, the data were further pre-processed to limit the transcriptome to a single transcript per gene. This step was conducted by using the *collapseRows* function of the WGCNA package in R (method = MaxMean), that identifies the transcript with fewest number of missing values and highest mean expression value across samples [12]. We limit the scope of this study to protein-coding genes with expression profiles summarized at the gene-level, in light of previously demonstrated drawbacks of isoform-level networks encompassing the non-coding transcriptome [13]. Inclusion of splice variants and non-coding RNAs increases network size by 100-fold [13], which in turn dramatically increases the computational resources necessary for network analysis. Although this high computational demand can theoretically be overcome by using a block-wise design where automatically selected subsets of the original data are used to build a co-expression network in a block-by-block fashion [1], previous work has shown that networks generated by block-wise design reflect only an approximation of networks generated by single-block design, which negatively influences the accuracy of subsequent module detection [13]. To avoid the shortcomings of a block-wise approach and pursue the most parsimonious hypothesis, our dataset included only protein-coding genes, with expression values summarized at the gene-level. This approach yielded a trimmed protein-coding gene expression matrix of 18,410 unique HGNC gene symbols.

To remove non-biological experimental variation, the dataset was adjusted for known batch effect (nucleic acid isolation batch) using the empirical Bayes framework employed by the *ComBat*

function of the Surrogate Variable Analysis (SVA) package in R [14] (as described in [15]). The gene expression data were further corrected for age, sex, death classification (assessed by the 4-point Hardy Scale) and post-mortem interval (PMI)-mediated covariance [16], added to the regression model as categorical or continuous variables as applicable (Fig. 1b, Supp. Fig. S1). The residuals calculated from this model were carried to downstream analysis. Finally, since low-expressed or non-varying genes usually represent noise in the data, we used the *goodSamplesGenes* function of the WGCNA package in R to iteratively identify and remove genes and samples with greater than 50% missing entries (default parameter) and genes with zero variance. The normalized, outlier-removed, residualized cortical expression values of 18,410 protein-coding genes from 107 samples constitute the final dataset for construction of the network.

#### *Network construction and module detection:*

The single-block pipeline implemented in the WGCNA R package was employed to build a signed and weighted gene co-expression network using the final dataset. Co-expression similarity was defined as the biweight midcorrelation (bicor) coefficient between the expression profiles of gene-pairs calculated for all possible comparisons. Bicor is a median-based measure of co-expression that was chosen as a robust alternative to mean-based similarity metrics (i.e., Pearson correlation) in evaluating similarity in gene expression [17]. To capture the continuous nature of pairwise interactions in biological systems and accentuate strong positive correlations, the resulting co-expression similarity matrix was transformed into a signed and weighted adjacency matrix. This step was conducted by using the soft-thresholding procedure implemented in the *pickSoftThreshold* and *scaleFreePlot* functions of the WGCNA package in R. A soft-thresholding power ( $\beta$ ) of 8 was identified as the lowest possible  $\beta$  yielding a power-law degree distribution that approximately fits one with a scale-free network topology, (signed  $R^2$  fit index = 0.8), while maintaining a relatively high mean connectivity (mean  $k > 100$ ) (Fig. 1c, Supp. Fig. S3). The rationale behind our choice to establish a scale-free topology is the demonstrated biological relevance of its degree distribution as a unifying property of many biological networks in nature, where a few hub nodes have many connections while most nodes have few connections [18-21]. Additionally, note that a demonstrated advantage of weighted correlation networks is the high robustness of the network construction to the choice of  $\beta$  parameter [22].

A signed adjacency matrix (Equation 1) was chosen over an unsigned adjacency matrix (Equation 2) to re-scale the underlying pairwise correlations from the  $[-1, 1]$  interval to the  $[0, 1]$  interval, as opposed to treating negative correlations as positive. In other words, since the adjacency matrix that underlies network construction is always non-negative, a signed network was chosen to respect the direction (up vs. down) of the co-expression relationships to prevent a clustering structure that mixes negatively correlated nodes, which often belong to different biological categories, with positively correlated nodes. The biological relevance of signed adjacency was demonstrated by previous work indicating that genes showing positive transcriptional correlation are more likely to exhibit known protein-protein interactions than uncorrelated or negatively correlated genes [10, 23, 24]. This approach is further supported by the recommendations of the creators of the WGCNA package in R (as described in <https://peterlangfelder.com/2018/11/25/signed-or-unsigned-which-network-type-is-preferable/>).

Signed adjacency  $A_{ij}^{signed}$  for genes  $i$  and  $j$  is defined as:

$$A_{ij}^{signed} = \left| \frac{1 + \text{bicor}(x_i, x_j)}{2} \right|^\beta \quad (1)$$

Unsigned adjacency  $A_{ij}^{unsigned}$  for genes  $i$  and  $j$  is defined as:

$$A_{ij}^{unsigned} = |\text{bicor}(x_i, x_j)|^\beta \quad (2)$$

The resulting signed and weighted adjacency matrix was transformed into a topological overlap (TO) matrix to capture not only the correlation between pairs of genes but also the connections among “neighborhoods” of genes [25, 26]. The TO measure (TOM) between two genes is high if the genes have many overlapping network connections, yielding a network interconnectedness measure that is proportional to the number of common neighbors shared between a pair of nodes. Several studies have shown that gene/protein-pairs with higher TO are more likely to play a role in the same functional class than gene/protein-pairs with lower TO [26-31], demonstrating that TOM yields biologically meaningful modules that can successfully capture the co-expression profile of genes encoding interacting proteins. Additionally, TOM was shown by previous work to be more robust to identifying spurious connections than pairwise correlation alone [32].

To explore the clustering structure of the nodes underlying our undirected network, we conducted average linkage hierarchical clustering on pairwise TOM, following its transformation into a dissimilarity measure ( $\text{dissTOM} = 1 - \text{TOM}$ ). To define network modules, we adaptively pruned the branches of the resulting dendrogram by using the dynamic-hybrid-tree-cut algorithm of the WGCNA package in R (Fig. 1d). This method employs a bottom-up approach that considers the shape of dendrogram branches in identifying clusters, yielding improved detection of outlying cluster members. This method was chosen, since it has been shown to outperform the traditional fixed-height branch cutting method for identifying biologically meaningful clusters, particularly in networks that exhibit a nested dendrogram structure [33]. Standard parameters were used to conduct this analysis: cut height = 99% of the truncated height range in the dendrogram, module detection sensitivity (deep split) = 2, minimum module size = 30, signed network with partitioning about medoids (PAM) respecting the dendrogram. To avoid over-clustering, we set the smallest number of genes that can be considered a module (minimum module size) to 30; this is a standard parameter used in the literature to establish a compromise between large modules that are robust and biologically informative and small modules that are possibly informative but less robust [34-36]. This approach yielded 31 modules, with module-sizes ranging from 43 to 3,319 genes; similar ranges have been reported in other applications.

We summarized the gene expression profiles of individual modules by eigengenes [1], identified as the first principal component of the expression data in a given module. Construction of eigengenes amounts to a network-based data reduction method that serves as a means to effectively correlate entire modules. Leveraging this approach, we amalgamated modules with very similar expression profiles to eliminate spurious assignment of highly co-expressed genes into separate modules. To this end, we conducted average linkage hierarchical clustering of module eigengenes (ME) on a correlation-based dissimilarity metric (1-pairwise Pearson's correlation between MEs) and merged modules that are strongly correlated ( $r > 0.8$  corresponding to cut height = 0.2) (Fig. 1d).

One of the resulting modules (grey module) was reserved for genes that could not be unequivocally assigned to any module (did not share similar co-expression patterns with the other genes of the network), as determined by the iterative refinement methodology used in our analysis pipeline to improve module detection. Therefore, the grey module was excluded from downstream analysis. Finally, we tested additional WGCNA parameters and determined that basic module structure in modules of interest (modules harboring 3q29 interval genes) was identifiable under variations of carefully selected algorithm parameters. A list of gene-sets for each identified module is provided in Table S1.2.

## Module preservation and quality analysis

To validate the reproducibility of our identified modules, we evaluated network preservation in an independent transcriptomic dataset, hereafter referred to as the test dataset/network. Publicly available RNA-Seq data from the BrainSpan Developmental Transcriptome Project [37] was used to build the test network (Supp. Fig. S4). 30 non-pathological post-mortem samples from the PFC of male and female adults (age range = 18-37) with no known neurological or psychiatric disorder were used in this study. To exclude any confounding pathology, specimens had been confirmed by neuropathological evaluation to not contain obvious malformations, extensive neuronal loss, neuronal swelling, or dysmorphic neurons and neurites. Additionally, any donor with a reported prolonged agonal condition (i.e., coma, hypoxia, seizures etc.), ingestion of neurotoxic substances at the time of death, suicide, severe head injury, significant hemorrhages, prominent vascular abnormalities, tumors, prominent brain lesions, stroke, congenital neural abnormalities, and signs of neurodegeneration (i.e., amyloid plaques, Lewy bodies etc.) were also excluded (see technical white paper [38]). Transcriptome profiling was performed using Illumina TruSeq RNA sequencing. Details on sample attributes and donor phenotypes are available in Table S1.1 and Supp. Fig. S4. The BrainSpan data (Developmental Transcriptome Dataset; v10 summarized to genes) used for the analyses described in this manuscript were downloaded from the Allen Brain Atlas portal (<https://www.brainspan.org/static/download/>) on 01/12/2019; the corresponding dbGaP study accession number is phs000755.v2.p1.

The test dataset was pre-processed following the same pipeline used for the GTEx reference dataset. We were unable to correct the normalized gene expression values of the test dataset for batch effect and death classification-mediated covariance, due to absence of the pertinent information from publicly available data. However, we note that the highly selective inclusion/exclusion criteria used by the BrainSpan project for tissue qualification [37] mitigates a potential confounding effect of death classification on gene expression. The inclusion of age, sex and PMI in the regression model was consistent with the pre-processing pipeline applied to the reference dataset. No outlier sample was identified by the above described ISC method. Notably, the test dataset was assembled by the aggregation of transcriptomic data from four sub-regions of the PFC (orbital, dorsolateral, ventrolateral, and medial PFC). To rule out a potential bias in test network construction driven by tissue-type, we conducted average linkage hierarchical clustering on the BrainSpan samples used in this study. The similarity metric for clustering was defined as the Pearson's correlation between gene expression levels of pairs of samples. The resulting dendrogram revealed no clustering pattern driven by tissue-type, ruling out PFC sub-region as a factor that could bias network construction (Supp. Fig. S4). Upon completion of pre-processing, the test dataset consisted of normalized and residualized gene expression values for 18,339 unique HGNC symbols from 30 samples.

To determine whether properties of within-module topology that were identified in our reference network were preserved in the test network, we calculated a composite network-based preservation statistic ( $Z_{\text{summary.pres}}$ ) [39] for each module by using the *modulePreservation* function of the WGCNA package in R.  $Z_{\text{summary.pres}}$  is a summary statistic that encompasses multiple density-based and connectivity-based preservation statistics that test 1) whether nodes sharing the same module in the reference network remain highly connected in the test network (i.e., are groups of genes defined as modules in the reference network denser than random groups of genes in the test network?) and 2) whether connectivity patterns between nodes underlying the reference network remain similar in the test network (i.e., do hub nodes of the reference network preserve their high degree of connectivity in the test network?). To determine whether the observed preservation statistics were higher than expected by chance and to derive a standardized Z score for each preservation statistic, we randomly permuted the module

assignments in the test data (number of permutations = 200; twice as high as default) and derived a  $Z_{\text{summary.pres}}$  score for each module. The resulting scores were evaluated according to established thresholds [39]:  $Z_{\text{summary.pres}} < 2$  indicates no evidence for preservation,  $2 < Z_{\text{summary.pres}} < 10$  indicates moderate evidence for preservation, and  $Z_{\text{summary.pres}} > 10$  indicates strong evidence for preservation. To account for the dependence of the  $Z_{\text{summary.pres}}$  score and permutation test p-values on module-size, we set the maximum module-size parameter of the *modulePreservation* function to 1000 genes (default parameter), reducing large modules by randomly sampling 1000 intra-modular nodes. Note that the specific goal of this preservation analysis was to determine the preservation strength of individual modules in relation to established  $Z_{\text{summary.pres}}$  score thresholds, as opposed to comparing the preservation statistics of modules with different sizes to one another (i.e., determine whether module A is more preserved than module B). Hence, aggregating multiple preservation statistics into an informative  $Z_{\text{summary.pres}}$  score constitutes a valid and advantageous approach for our purposes, despite its sensitivity to module-size.

In addition to measuring the density and connectivity-based preservation of each module between the reference and test networks, we measured the quality of the identified modules in the reference network without a reference to the test network. The goal of this analysis was to assess the robustness of the identified modules (i.e., how distinct is a given module from other modules in the reference network?) by calculating a composite quality statistic ( $Z_{\text{summary.qual}}$ ) [39] for each module, as implemented in the *modulePreservation* function of the WGCNA package in R. This approach is akin to a cluster stability analysis [40] and employs a resampling technique that applies the module preservation statistics outlined above to repeated random splits of the reference data. The resulting  $Z_{\text{summary.qual}}$  score indicates the robustness of a given module definition (hence the parameters selected for network construction and module detection) across networks created from the original reference data. The same  $Z_{\text{summary.pres}}$  thresholds outlined above were used to evaluate the  $Z_{\text{summary.qual}}$  scores. Finally, in line with the recommendations [39] of the creators of the WGCNA package, we also evaluated the individual preservation and quality statistics underlying the composite scores for each module (Table S1.4, Supp. Fig. S5). Results from the module preservation and quality analysis (Fig. 1e) revealed the replicable and robust nature of the network, thus we commenced interrogation of the network for insights into 3q29 interval gene function.

### Functional characterization of network modules harboring 3q29 genes

Understanding the functions of the 21 protein-coding genes hemizygotously deleted in 3q29Del syndrome is an obligatory step towards gaining insights into the cellular mechanisms underlying disease etiology. Others have shown that each gene of the human genome is estimated on average to be involved in ten biological functions [41]. Given that many of the genes in the 3q29Del interval are poorly annotated and some lack a functional annotation altogether (Supp. Fig. S6), we leveraged the transcriptomic co-expression approach to predict gene function. We screened each network module for membership of 3q29 interval genes and exploited major biological databases to derive a functional interpretation of the co-expression modules found to harbor at least one 3q29 interval gene (hereafter referred to as a 3q29 module).

Functional enrichment analyses of individual 3q29 modules (transformed into gene-sets) were run on the g:Profiler webserver (<http://biit.cs.ut.ee/gprofiler>; Ensembl version 96, Ensembl Genomes version 43) by using gene ontology biological processes (GO:BP), and Reactome (REACT) and Kyoto Encyclopedia of Genes and Genomes (KEGG) biological pathways. For high-confidence analysis, electronic GO annotations ("IEA"), which are assigned to gene products without curator verification (majority inferred *in silico*) and often cannot be traced to an experimental source, were discarded; filtering GO annotations based on evidence code has been

recommended in previous literature to avoid erroneous results [42]. The statistical domain scope for functional enrichment analysis was set to only genes with at least one known annotation (default parameter) to establish an effective genomic background for statistical testing using the hypergeometric probability function. Enriched terms surpassing the adjusted g:SCS significance threshold of  $P < 0.05$  were filtered for gene-set size (allowed range: 10-2,000 genes) and semantic similarity to improve the specificity and interpretability of our results [43]. The g:SCS (Set Counts and Sizes) method for multiple comparisons correction was our method of choice for pathway analysis, since it was shown to outperform standard approaches, such as Bonferroni correction, in estimating the true effect of multiple testing over the complex structure of functional profiling data [44]. Functional classifications in databases such as gene ontology (GO) have a heavily overlapping hierarchical structure, since any term is automatically related to all other terms included in its relational path. Hence, statistical assumptions underlying more traditional correction methods such as Bonferroni, which was designed for multiple independent tests, cannot be met in our application. Hence, we used the novel g:SCS method (refer to [43] for details) as an alternative solution to the multiple testing problem that is highly complex in functional data. Note that the g:Profiler software that was used to perform pathway analyses in this study also implements the g:SCS significance thresholds by default. The REVIGO webtool (<http://revigo.irb.hr/>) was used to reduce redundancy in the identified GO terms (default semantic similarity measure = SimRel, allowed similarity = 0.9 (large)) [45]. Top 10 biological pathways (REACT) found to be enriched in individual 3q29 modules were ranked by statistical significance level (adjusted  $P < 0.05$ ) and are shown in the main results. Extended results are provided in Table S1.5.

Furthermore, findings of the functional enrichment analysis were also evaluated to determine whether identified co-expression modules reflect true biological signals as opposed to noise; this distinction was inferred by enrichment of the constituent genes for known biological processes and pathways. Our gene ontology-based approach complemented the module quality analysis described above.

To further interrogate whether the gene co-expression modules identified in this study represent biologically meaningful units of nodes with shared membership of the same molecular complex or functional pathway, we also investigated whether the genes co-clustering in the same transcriptomic module tend to interact at the protein level. First, we queried the known protein interactors of 3q29 interval genes identified by the Human Reference Protein Interactome Mapping Project (HuRI; <http://interactome-atlas.org/>). The interaction data in HuRI were retrieved from two sources: 1) a systematic binary mapping pipeline using a high-throughput yeast two-hybrid assay, followed by retesting and validation, and 2) interactions curated from the literature and further filtered by HuRI to identify high-quality binary interactions (see details in [46]).

Second, we tested the co-expression modules harboring 3q29 interval genes for enrichment of known and predicted protein-protein interactions (PPIs) from the STRING database (v.11, <https://string-db.org/>). Using the STRING search tool, PPIs were retrieved from active interaction sources (species: "Homo sapiens"), including experiments (biochemical data), databases (previously curated pathway and protein-complex knowledge), genomic context prediction channels (neighborhood, fusion, gene co-occurrence), and text mining. Interactions that rely on STRING's RNA-seq co-expression inference pipeline were excluded from this analysis. A minimum interaction confidence score of 0.4 (medium confidence, default setting) was applied to construct a PPI network for each module; this interaction score represents the approximate confidence of an association based on all available evidence from active sources. The STRING enrichment analysis tool was used to test whether the observed number of interactions (edges) in each interrogated module was significantly higher than the number of edges expected if the

nodes were to be selected from the genome at random (see details in [47]). STRING enforces an upper limit on the number of query items and does not support the PPI network enrichment analysis of > 2,000 nodes. Hence, 2,000 genes from the green and turquoise modules, which harbor 4,746 and 3,319 genes respectively, were randomly subsampled without replacement, using the *sample* function in R (v. 4.0.3). The 3q29 interval genes clustering in each of these two modules were used as forced entries during this subsampling procedure to ensure that each resulting PPI subnetwork included our primary genes of interest.

### Identification of meta-modules harboring 3q29 genes

Several studies have demonstrated that individual modules can be organized into biologically meaningful meta-modules that represent a higher-order organization of the transcriptome [48], likely reflecting pathway dependencies and synergistic function. To evaluate whether individual 3q29 modules clustered together within larger meta-modules of the co-expression network, we investigated the relationship among modules by leveraging their eigengenes (ME). We calculated the Pearson's correlations between all pairs of MEs and used this similarity metric to conduct average linkage hierarchical clustering of the MEs. Consistent with its use in other applications, we defined meta-modules as tight clusters of positively correlated MEs detectable as major branches of the resulting eigengene dendrogram [48]. The identified meta-modules were screened to identify the grouping patterns among 3q29 modules across the co-expression network.

This eigengene-based network reduction framework was further utilized to determine whether individual 3q29 modules that were found to partake in larger meta-modules were truly distinct from each other. To this end, gene ontology information was leveraged to identify common versus distinct enrichment of biological processes and pathways in individual 3q29 modules sharing a meta-module. This approach also complemented the module quality analysis described above.

### Determination of module membership strength for 3q29 genes

To measure how strongly connected individual 3q29 genes are to their assigned modules, we used the *signedKME* function of the WGCNA package in R to calculate a module membership measure (kME) for each 3q29 gene. kME is an eigengene-based module connectivity measure, defined as the Pearson's correlation between the expression profile of a given gene and the eigengene (first principal component) of a given module[1]. An advantage of using kME as a network connectivity metric is that it allows the direct comparison of module membership values across modules that differ in size. This property distinguishes kME from degree-based measures of module connectivity, which are derived by summing a node's total number and strength of connections within a module and result in a metric that can be biased by module size.

A kME of 0 indicates that a gene is uncorrelated with an ME of interest and is thus unlikely to be a member of that module. In contrast,  $kME > 0.8$  describes a hub gene that is highly connected to other genes in its module, hence predicted to be a crucial component of the overall function of that module. Previous work has shown that targeted disruption of a hub gene has a more deleterious effect on the ability of the network to function and leads to a larger number of phenotypic outcomes than the disruption of randomly selected genes or targeted deletion of less connected genes [49, 50]. These observations have led to the hypothesis that hub nodes of biological networks are typically associated with human disease genes. Indeed, several lines of evidence, particularly from cancer biology, have validated this hypothesis [51, 52]. One study has shown that topological features of disease-genes identified from OMIM's Morbid Map of the

Human Genome disproportionately exhibit hub-gene characteristics, with protein products participating in more known protein-protein interactions than that of non-disease genes[53]

Note that there is no established definition of a hub node in network analysis, since the selection criteria can vary depending on the sparsity of the network; examples of hub gene definitions based on  $kME \geq 0.7$  exist in the literature. To generate rigorous WGCNA-based predictions, we adopted a conservative criterion that defines hub genes as nodes with  $kME > 0.8$  ( $P < 0.05$ ), a stringent threshold used by several other studies [54-56]. We annotated 3q29 interval genes that surpassed this kME threshold as hub genes, whose loss of function is predicted to produce a highly deleterious impact on the system.

Additionally, we screened kME values to identify 3q29 genes with very weak module membership. Given that kME quantifies how close a gene is to its assigned cluster of co-expressed genes, we excluded 3q29 genes with non-significant kMEs ( $P > 0.05$ ) from downstream analysis. Notably, we limited our evaluations to intra-modular kME values that describe a gene's correlation with the eigengene of its assigned module. In rare instances, the kME of a given gene was found to be slightly higher for a module other than its assigned module. This finding stems from the fact that TOM yields a measure of network interconnectedness that is similar but not identical to an only correlation-based approach (as described by the creators of the WGCNA package at <https://support.bioconductor.org/p/101579/>). As noted previously, TOM-based similarity was shown to outperform correlation-based similarity in identifying biologically meaningful modules [32]. A list of kME values and corresponding p-values computed via correlation tests is provided in Table S1.3 for each node and module.

### Identification of prioritized driver genes

3q29Del confers >40-fold increased risk for SZ and constitutes a shared risk factor for IDD and ASD. However, no single gene in this interval has been definitively associated with SZ, IDD, or ASD. To generate testable hypotheses about which 3q29 interval genes are causally linked to the major neuropsychiatric phenotypes associated with 3q29Del, we further leveraged our gene co-expression network. We adopted the guilt-by-association approach and screened the intra-modular subnetworks of individual 3q29 interval genes for a significant overlap with known SZ, ASD and IDD-risk genes. Guilt-by-association is a widely used principle that operates on the assumption that disease-associated genes are more closely connected to each other than random pairs of nodes in a network. Ample evidence demonstrates the high utility of the guilt-by-association approach in identifying novel disease risk genes [57-61].

Previous work has shown that testing the overrepresentation of known disease-risk genes in entire network modules harboring a large number of genes produces an inflated false positive rate[13]. To avoid this, we reduced 3q29 modules to top-neighbor-based subnetworks, maintaining only the close intra-modular connections of individual 3q29 genes. A top neighbor was defined as any node whose gene expression profile has a moderate-to-high correlation (Spearman's  $\rho$  ( $p$ )  $\geq 0.5$ ) with a given 3q29 interval gene (considered a "seed" node) within the same module. Note that top neighbors were identified by a correlation-based hard-thresholding method applied only to intra-modular edges connecting seed-node pairs. Hence, the criterion for top neighbor identification combines the strength of scale free-topology and topological overlap principles of WGCNA for network construction and module detection, with subsequent application of a selective hard-thresholding method, resulting in a binary classification of top neighbors predicted to form direct functional links with 3q29 interval genes.

We conducted hypergeometric tests to determine whether top-neighbor-based intra-modular subnetworks of individual 3q29 genes are enriched for curated gene-sets with known SZ, ASD or IDD association. Six gene-sets were curated for this purpose from the following sources: 1) 93 IDD-risk genes enriched for damaging *de novo* mutations, identified by the Deciphering Developmental Disorders Study [62]; 2) 86 ASD-risk genes categorized by the Simons Foundation Autism Research Initiative (SFARI) [63, 64] as “high-confidence candidate risk genes” with strong evidence for ASD association (categories = 1 & 2; downloaded on 2/15/2019 from <https://gene.sfari.org/database/>); 3) 651 SZ-related genes shown by Meng et al. (2018) [65] to demonstrate significant differential expression in the postmortem brain tissue of SZ cases/controls, identified from the PsychENCODE BrainGVEX dataset [66]; 4) 636 SZ-related genes shown by Fromer et al. (2016) to demonstrate significant differential expression in the postmortem dorsolateral PFC tissue of SZ cases/controls, identified from the CommonMind Consortium dataset [67]; 5) 340 SZ-risk genes adjacent to SZ-associated genetic loci, identified by the most recent genome-wide association study (GWAS) [68] conducted by the Psychiatric Genomics Consortium (PGC); 6) 290 SZ-risk genes with exonic *de novo* mutations, identified via the Neuropsychiatric Disorder *De Novo* Mutations Database [69] (downloaded on 2/17/2019 from <http://www.wzgenomics.cn/NPdenovo/>). Notably, these gene-sets were selectively curated from the literature to obtain a comprehensive yet reliable list of reported IDD, ASD and SZ-associated genetic variants spanning a wide range of the allele frequency spectrum. As noted previously, 3q29 genes with non-significant intra-modular kMEs ( $P > 0.05$ ) (Fig. 2c) were excluded from this analysis.

To evaluate the specificity of the investigated disease enrichment patterns, we also tested the significance of the overlap between 3q29 subnetworks and negative control gene-sets associated with Parkinson’s disease (PD), late-onset Alzheimer’s disease (AD) and inflammatory bowel disease (IBD). The gene lists for these conditions were considered “negative controls,” as they constitute disease phenotypes (two related and one unrelated to brain health) with no known association to 3q29Del. To rule out a potential bias that could be introduced to our enrichment analysis by differences in the sizes of curated gene-sets, common genetic variants associated with height (large gene-set size comparable to several SZ gene-set sizes) were included as another negative control. Four negative control gene-sets were curated for this purpose from the following sources: 1) 25 AD-risk genes identified by the largest published GWAS meta-analysis conducted by the Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for AD Consortium (GERAD/PERADES) [70]; 2) 67 PD-risk genes identified by the largest published GWAS meta-analysis conducted by the International PD Genomics Consortium [71]; 3) 98 IBD-related genes identified by the International IBD Genetics Consortium as “strong positional candidate genes” in GWAS-identified risk loci [72]; 4) 479 height-associated genes identified by the largest published GWAS meta-analysis conducted by the Genetic Investigation of Anthropometric Traits Consortium (GIANT) [73].

To accurately measure the union-size of the possible matches between the curated gene-sets and WGCNA-derived 3q29 subnetworks, genes that could only be present in one set (non-protein-coding genes; gene-type annotated by GENCODE v19) were excluded from the overlap analysis, yielding the final gene-set sizes listed above. Notably, discrepancies in gene symbol alias usage were accounted for to ensure consistency in nomenclature. The background list for the hypergeometric test was determined as the total number of unique genes used to conduct WGCNA. The *GeneOverlap* package [74] in R (version 1.20.0) was used for this analysis. The hypergeometric p-values obtained by overlap analysis were corrected for multiple testing using the Benjamini-Hochberg procedure ( $n = 10$  gene-sets). 3q29 genes whose subnetworks were found to show a significant overlap with SZ, ASD and/or IDD risk genes (adjusted  $P < 0.05$ ) were prioritized as driver genes, along with their SZ, ASD, and/or IDD-related top neighbors from the

corresponding enriched disease gene-set. These prioritized driver genes are predicted to contribute to the emergence of the major neuropsychiatric phenotypes associated with 3q29Del. The list of curated gene-sets, WGCNA-derived top neighbors, WGCNA-derived prioritized driver genes and detailed results of the overlap analysis are provided in Tables S2.1-3.

### **Functional characterization of prioritized driver genes**

According to the “local hypothesis” proposed by the emerging paradigm of network medicine, the human transcriptome and proteome demonstrate non-random topological characteristics, where disease genes tend to interact with other disease genes and play distinct roles in disrupting the same biochemical process underlying a common pathophenotype [75]. Motivated by this theory, we sought to formulate testable hypotheses about the key biological mechanisms linking 3q29Del to SZ, ASD, and IDD by conducting a functional enrichment analysis on the union of our prioritized driver genes. We used the same analysis approach described above for testing the functional enrichment of 3q29 modules. The biological processes and pathways that were found to surpass the adjusted g:SCS significance threshold of  $p < 0.05$  were filtered for gene-set size and semantic similarity; the top 20 biological processes and pathways found to be enriched in our prioritized driver genes are shown in the main results. To provide a thorough illustration of all enriched GO:BP terms in the main results, GO:BP findings were further organized into a network visualization of related functional annotation categories. Extended results are provided in Table S2.5.

Overall, the transcriptomic network identified in this study is predicted to connect 3q29 interval genes with gene-sets outside the interval that participate in the same or overlapping biological process and associate with similar disease phenotypes. Perturbation of 3q29 interval gene dosage is expected to also perturb the functioning of network-partners outside the recurrent 3q29Del locus. However, note that the underlying structure of weighted gene co-expression networks is agnostic to the mechanistic order of cellular and molecular events. The information necessary to derive the order of biological interactions is not an explicit outcome of gene co-expression itself, since such inferences require time-dependent analysis of combinatorial interactions between nodes. As a result, some of the network partners identified in this study are expected to function upstream of their 3q29 gene partner and would likely not be affected by 3q29Del.

### **Proof of concept study for testing the validity of WGCNA-based predictions.**

#### *Overview:*

A necessary step in determining the utility of network-based predictions is a proof of concept of their validity in an experimental system. To this end, we assessed the validity of our WGCNA-derived predictions by testing the enrichment of the identified network-partners of 3q29 interval genes for differential expression in the mouse model of 3q29Del [76].

#### *RNA-sequencing in mouse cortex*

Mice harboring a heterozygous deletion of 1.26Mb (Del16<sup>+/Bdh1-Tfrc</sup>) that is homologous to the human 3q29Del locus were generated by CRISPR/Cas9 technology previously [76]. At postnatal day seven, five mutant and five wild-type male pups were anesthetized under isoflurane and rapidly decapitated. The bilateral cortical sheet was dissected, chopped with a scalpel, and homogenized in QIAzol (Qiagen) in a Bullet Blender Tissue Homogenizer (Next Advance, Inc., Troy, NY). Total RNA was isolated using the miRNeasy Mini Kit (Qiagen) with on-column DNase I treatment (Qiagen). Sequencing libraries were generated using the SMART-Seq Stranded Kit (Takara Bio, Mountain View, CA). 50M paired-end 150bp read sequencing was performed on an

Illumina platform. Sequences were quality-checked and aligned to the mm10 reference genome. Gene quantification was conducted using HTSeq-count [77].

*Differential gene expression analysis:*

We used two analysis tools (DESeq2 [78] and edgeR [79]) using the negative binomial model to identify differentially expressed genes (DEGs). Since there is no established consensus on a gold-standard statistical pipeline for conducting differential expression analysis in transcriptomic research, we incorporated both programs into our analysis as two independent methodologies representing the state of the art in bioinformatics [80]. Only the protein-coding consensus DEGs that were concurrently identified by DESeq2 (version 1.24.0) and edgeR (version 3.26.8) were carried into downstream analysis.

Read counts from technical replicates were summed up using the *collapseReplicates* and *sumTechReps* functions of the DESeq2 and edgeR packages, respectively. This approach effectively increases the sequencing depth of the individual biological replicates, thereby increasing the power to detect differential expression. Genes with low counts were filtered by mean normalized counts in DESeq2 and by expression in counts per million (CPM) in edgeR. To minimize false negatives, we cast a wide net and determined statistically significant differences in gene expression at the nominal significance level ( $P < 0.05$ ); a similar approach has been taken in previous transcriptomic studies modelling neuropsychiatric disorders [81, 82].

*Comparison of empirically identified DEGs and network-derived predictions:*

We tested the statistical significance of the overlap identified between DEGs found in the mouse model of 3q29Del and the co-expression partners of 3q29 interval genes identified by WGCNA via hypergeometric tests. A similar proof of concept approach has been used in previous literature [65]. The *GeneOverlap* package [74] in R was used for this analysis. We investigated intersections at three scales of network interconnectedness: i) the broad 3q29 network, ii) the top-neighbor-based 3q29 subnetwork, and iii) the prioritized driver genes. To accurately measure the union-size of the possible matches between mouse DEGs and WGCNA-based targets, all compared gene-sets were filtered for known human-mouse homologs as determined by the HomoloGene database of the National Center for Biotechnology Information (NCBI) [83], using the *homologene* package (version 1.4.68.19.3.27) in R.

i) The broad 3q29 network is comprised of the union of seven WGCNA-derived modules that were found to harbor at least one 3q29 interval gene. The constituent genes of 3q29 modules show high topological overlap with one or more 3q29 genes and with one another, forming tight clusters of nodes that not only show high pair-wise co-expression with one another but also share many network neighbors. The clustering structure derived from such coordinated expression patterns in local regions of the genome likely reflects co-regulation and/or shared function among constituting genes. Hence, the union of the 3q29 modules identified in this study represents a broad subset of the human protein-coding genome that shows coordinated expression at the mRNA level with the 21 protein-coding genes located in the interval. A total of 11,924 genes with known human-mouse homology, including 21 3q29 interval genes comprise this broad network.

ii) The top-neighbor-based 3q29 subnetwork represents a refined subgraph, where modules are restricted to only the “top neighbors” of 3q29 interval genes. These top neighbors are predicted to function as direct interacting partners of 3q29 genes, participating in the same or overlapping biological pathways within the modular organization of molecular systems. A top neighbor was defined as any node whose gene expression profile has a moderate-to-high pairwise correlation ( $\rho \geq 0.5$ ,  $P < 0.05$ ) with a 3q29 interval gene (considered a “seed” node) within the same module. The network ties underlying this subnetwork were derived by a correlation-based hard-

thresholding method applied only to intra-modular edges connecting seed-node pairs. Hence, the top neighbor criterion used to construct this sub-network combines the strengths of scale free-topology and topological overlap principles for network construction and module detection, with subsequent application of a hard-thresholding method. The total number of genes in this subnetwork with known human-mouse homology is 5,087, including 21 3q29 interval genes.

iii) The prioritized driver genes constitute the most refined subset of the transcriptomic network connections identified in this study. These are comprised of select 3q29 genes and top neighbors that are predicted to function as the primary drivers of the neurodevelopmental and psychiatric consequences of 3q29Del. These drivers were identified by leveraging the widely used guilt-by-association principle predicated on the assumption that disease-associated genes are more closely connected to each other than random pairs of nodes in a network. As described in previous sections of our methods, we conducted hypergeometric tests to determine whether the top neighbors of individual 3q29 genes are enriched for curated gene-sets with known SZ, ASD or IDD association. 3q29 genes whose top neighbors were found to show a significant overlap with known risk genes (adjusted  $P < 0.05$ ) were prioritized as driver genes, along with their disease-related top neighbors from the corresponding enrichment analyses. The total number of prioritized driver genes with known human-mouse homology is 280, including nine 3q29 interval genes.

Note that the underlying structure of weighted gene co-expression networks is agnostic to the mechanistic order of cellular and molecular events. As a result, some of the network targets identified in this study are expected to function upstream of their 3q29 gene partner and would likely not be affected in the mouse model of 3q29Del.

## **Extended Results**

### **Unbiased gene co-expression network analysis reveals convergent and distributed effects of 3q29 interval genes across the adult human cortical transcriptome.**

Our WGCNA-based unsupervised network analysis approach, applied to publicly available high-throughput GTEx data, revealed that the protein-coding transcriptome of the healthy adult human PFC can be organized into a co-expression network of 19 modules (labeled by color) (Fig. 1d, Fig. 2a). The identified modules group genes with highly similar expression profiles into densely interconnected clusters, which likely represent shared function and co-regulation. One of the identified modules (the grey module) contained genes that could not be unequivocally assigned to any module; thus, it was excluded from downstream analysis. The resulting module sizes (number of genes assigned to a module) ranged from 43 (steel blue) to 4,746 (green) genes, with an average module size of 1,014 genes (excluding grey module). Similar ranges have been reported in other network analysis applications. A list of gene-sets for each identified module is provided in Table S1.2.

The 21 protein-coding genes located in the 3q29 interval were found to cluster into seven network modules, which represent local regions of the human protein-coding genome that demonstrate coordinated expression with the 3q29 locus (Fig. 2a). These modules are referred to as 3q29 modules, and were labeled as black (size = 1,170 genes), brown (size = 1,972 genes), dark turquoise (size = 496 genes), green (size = 4,746 genes), magenta (size = 1,437 genes), midnight blue (size = 1,414 genes), and turquoise (size = 3,319 genes) modules. Moreover, 18 / 21 3q29 interval genes were found to concentrate into just four modules (brown, green, midnight blue, turquoise) (Fig. 2a), suggesting that the haploinsufficiency of the 3q29 locus may perturb the

same biological processes via multiple hits, cumulatively disrupting redundancy and compensatory resiliency in the normative regulation of cellular functions. Simultaneously, leading candidate genes *DLG1* (black) and *PAK2* (dark turquoise), which were previously hypothesized to contribute to neuropsychiatric pathology, were found in opposite branches of the network, demonstrating the potential distributed effects of this CNV across the transcriptomic landscape (Fig. 2a). The modular allocation of 3q29 interval genes is listed below:

<b>Modules harboring 3q29 interval genes (“3q29 Modules”)</b>		
<b>Network module</b>	<b>Number of 3q29 interval genes</b>	<b>Constituent 3q29 interval gene symbols</b>
Black	1	<i>DLG1</i>
Brown	4	<i>NCBP2</i> , <i>TFRC</i> , <i>TM4SF19</i> , <i>ZDHHC19</i>
Dark turquoise	1	<i>PAK2</i>
Green	6	<i>BDH1</i> , <i>PIGZ</i> , <i>PCYT1A</i> , <i>SMCO1</i> , <i>SLC51A</i> , <i>MFI2</i>
Magenta	1	<i>RNF168</i>
Midnight blue	3	<i>UBXN7</i> , <i>SENP5</i> , <i>WDR53</i>
Turquoise	5	<i>CEP19</i> , <i>FBXO45</i> , <i>PIGX</i> , <i>TCTEX1D2</i> , <i>NRROS</i>

The average linkage hierarchical clustering of the module eigengenes revealed that the identified modules further clustered into three higher level meta-modules (clusters of highly correlated modules), detected as major branches of the resulting eigengene dendrogram (Fig. 2a, 2b). The magenta, green and turquoise 3q29 modules clustered together within the first meta-module, grouping 12 3q29 interval genes: *RNF168*, *BDH1*, *PIGZ*, *PCYT1A*, *SMCO1*, *SLC51A*, *MFI2*, *CEP19*, *FBXO45*, *PIGX*, *TCTEX1D2* and *NRROS*. The brown and dark turquoise 3q29 modules clustered together within a second meta-module, grouping five 3q29 interval genes: *NCBP2*, *TFRC*, *TM4SF19*, *ZDHHC19* and *PAK2*. Finally, the midnight blue and black 3q29 modules were found to cluster together within a third meta-module, grouping three 3q29 interval genes: *DLG1*, *UBXN7*, *SENP5* and *WDR53*. The observed grouping, as well as the segregation, of sets of 3q29 modules into distinct meta-modules represents a higher-order transcriptomic organization of the 3q29 locus, which likely reflects pathway dependencies and interactions between biological processes involving 3q29 interval genes. These findings suggest that, rather than functioning as independent non-interacting units, sets of 3q29 interval genes and their co-expressed network partners may work in synergy at both the module and meta-module levels of transcriptomic organization (Fig. 2a, 2b), and likely constitute interacting sources of pathology in 3q29Del syndrome.

### **Pathway analysis points to functional involvement of the 3q29 locus in nervous-system functions and core aspects of cell biology.**

The functional enrichment analysis of individual 3q29 modules showed that the constituent genes of each module load highly onto canonical biological processes and pathways (Fig. 2d). These functional enrichment findings validate that our co-expression-based 3q29 modules reflect clustering dynamics that are biologically meaningful.

Functional characterization of the black module: We observed that the biological processes and pathways that were significantly overrepresented in the black module mainly encompass terms related to regulation of gene expression and maintenance of the integrity of the cellular genome. Based on the ranking of p-values adjusted for multiple testing, the top biological pathways (annotated by the Reactome database) that were overrepresented in the black module include metabolism of RNA (REAC:R-HSA-8953854, adjusted  $P = 2.45E-07$ ), processing of capped intron-containing pre-mRNA (REAC:R-HSA-72203, adjusted  $P = 7.74E-04$ ), chromatin

organization (REAC:R-HSA-4839726, adjusted  $P = 6.36E-03$ ), mRNA splicing (REAC:R-HSA-72172, adjusted  $P = 2.11E-02$ ), post-translational protein modification (REAC:R-HSA-597592, adjusted  $P = 2.33E-02$ ), DNA Repair (REAC:R-HSA-73894, adjusted  $P = 4.02E-02$ ) and tRNA processing (REAC:R-HSA-72306, adjusted  $P = 4.97E-02$ ).

Functional characterization of the midnight-blue module: Similarly, the midnight-blue module, which is in the same meta-module as the black module, was found to be enriched for biological pathways and processes that are involved in DNA repair and regulation of gene expression at the levels of transcription and translation, as well as cellular response to stress. An important functional signature that set the midnight blue module apart from the black module was its specific enrichment for terms related to cell cycle regulation. The top biological pathways that were overrepresented in the midnight blue module include gene expression (transcription) (REAC:R-HSA-74160, adjusted  $P = 2.10E-39$ ), metabolism of RNA (REAC:R-HSA-8953854, adjusted  $P = 1.64E-06$ ), DNA double-strand break repair (REAC:R-HSA-5693532, adjusted  $P = 9.79E-04$ ), mRNA 3'-end processing (REAC:R-HSA-72187, adjusted  $P = 5.49E-04$ ), cell cycle (REAC:R-HSA-1640170, adjusted  $P = 3.26E-06$ ), mRNA splicing (major pathway) (REAC:R-HSA-72163, adjusted  $P = 5.33E-03$ ), cell cycle checkpoints (REAC:R-HSA-69620, adjusted  $P = 5.73E-03$ ), transport of mature mRNA derived from an intron-containing transcript (REAC:R-HSA-159236, adjusted  $P = 1.53E-02$ ), and transcriptional regulation by the tumor suppressor TP53 (REAC:R-HSA-3700989, adjusted  $P = 2.10E-02$ ).

Functional characterization of the brown module: Assessment of shared function among constituent genes of the brown module revealed primary enrichment for biological pathways and processes involved in cellular metabolism and mitochondrial function. The top biological pathways that were overrepresented in the brown module include fatty acid metabolism (REAC:R-HSA-8978868, adjusted  $P = 7.59E-08$ ), mitochondrial fatty acid beta-oxidation (REAC:R-HSA-77289, adjusted  $P = 8.33E-05$ ), chondroitin sulfate/dermatan sulfate metabolism (REAC:R-HSA-1793185, adjusted  $P = 1.42E-03$ ), peroxisomal protein import (REAC:R-HSA-9033241, adjusted  $P = 1.79E-03$ ), metabolism of fat-soluble vitamins (REAC:R-HSA-6806667, adjusted  $P = 4.99E-03$ ), peptide hormone biosynthesis (REAC:R-HSA-209952, adjusted  $P = 8.72E-03$ ), solute-carrier (SLC)-mediated transmembrane transport (REAC:R-HSA-425407, adjusted  $P = 8.94E-03$ ), and diseases associated with glycosaminoglycan metabolism (REAC:R-HSA-3560782, adjusted  $P = 1.30E-02$ ). Notably, the brown module was also found to be enriched for two canonical KEGG-annotated pathways: the Hippo signaling pathway (KEGG:04390, adjusted  $P = 4.82E-03$ ) and the Wnt signaling pathway (KEGG:04310, adjusted  $P = 4.96E-02$ ), which play crucial roles in growth and developmental pathways with substantial cross-talk [84].

Functional characterization of the dark turquoise module: The dark turquoise module was found to coalesce genes that are enriched for biological functions in epigenetic regulation of gene expression, as well as in signal transduction pathways that are mediated by Rho GTPases. This latter function is at least in part attributable to *PAK2*, the only 3q29 interval gene assigned to this module, which encodes a known Rho GTPase effector. Intriguingly, this module was also found to be enriched for a functional role in estrogen receptor-mediated signaling. In particular, the top terms that were overrepresented in the dark turquoise module include estrogen-dependent gene expression (REAC:R-HSA-9018519, adjusted  $P = 5.08E-06$ ), estrogen receptor (ESR)-mediated signaling (REAC:R-HSA-8939211, adjusted  $P = 9.52E-06$ ), Rho GTPase effectors (REAC:R-HSA-195258, adjusted  $P = 1.35E-05$ ), SIRT1 negatively regulates rRNA expression (REAC:R-HSA-427359, adjusted  $P = 2.09E-05$ ), gene silencing by RNA (REAC:R-HSA-211000, adjusted  $P = 3.95E-05$ ), nucleosome assembly (REAC:R-HSA-774815, adjusted  $P = 4.75E-05$ ), chromatin modifying enzymes (REAC:R-HSA-3247509, adjusted  $P = 5.12E-05$ ), signaling by nuclear receptors (REAC:R-HSA-9006931, adjusted  $P = 8.46E-05$ ), meiotic synapsis (REAC:R-HSA-

1221632, adjusted  $P = 8.73\text{E-}05$ ), and epigenetic regulation of gene expression (REAC:R-HSA-212165, adjusted  $P = 1.44\text{E-}04$ ). Notably, the dark turquoise module was found to share the same meta-module as the metabolism-related brown module, suggesting the involvement of several 3q29 interval genes in a hierarchical transcriptomic control structure that interconnects Rho GTPase-mediated signaling cascades and estrogen-regulated signal transduction pathways with metabolic regulation. Emerging empirical findings demonstrate the existence of a crosstalk between these fundamental pathways [85-87], supporting the biological relevance of the co-expression-based clustering patterns underlying the meta-module that harbors the brown and dark turquoise modules identified in this study.

Functional characterization of the turquoise module: The biological processes that were overrepresented in the turquoise module primarily encompass nervous-system specific terms comprising the regulation of nervous system development and function. Other enriched biological functions that are non-specific but cardinal to nervous-system operations involve ion transport, calcium signaling, cyclic adenosine monophosphate (cAMP)-dependent signal transduction and cell projection organization. Specifically, the top Reactome-based biological pathways that were enriched in the turquoise module include neuronal system (REAC:R-HSA-112316, adjusted  $P = 2.98\text{E-}12$ ), protein-protein interactions at synapses (REAC:R-HSA-6794362, adjusted  $P = 7.12\text{E-}07$ ), neurexins and neuroligins (REAC:R-HSA-6794361, adjusted  $P = 3.58\text{E-}06$ ), transmission across chemical synapses (REAC:R-HSA-112315, adjusted  $P = 6.31\text{E-}06$ ), neurotransmitter receptors and postsynaptic signal transmission (REAC:R-HSA-112314, adjusted  $P = 9.47\text{E-}05$ ), serotonin receptors (REAC:R-HSA-390666, adjusted  $P = 6.42\text{E-}03$ ), the citric acid (TCA) cycle and respiratory electron transport (REAC:R-HSA-1428517, adjusted  $P = 7.00\text{E-}03$ ), and unblocking of N-methyl-D-aspartate (NMDA) receptors, glutamate binding and activation (REAC:R-HAS 438066, adjusted  $P = 1.41\text{E-}02$ ). Complementing these biological pathways, GO biological processes that were found to be enriched in the turquoise module include regulation of synaptic plasticity (GO:0048167, adjusted  $P = 6.68\text{E-}03$ ), cognition (GO:0050890, adjusted  $P = 2.91\text{E-}02$ ), neuron differentiation (GO:0030182, adjusted  $P = 3.19\text{E-}02$ ), long-term potentiation (GO:0060291, adjusted  $P = 3.81\text{E-}02$ ) and learning and memory (GO:0007611, adjusted  $P = 4.20\text{E-}02$ ). Dysregulation of these biological processes and pathways has been implicated in the etiology of major neuropsychiatric and neurodevelopmental disorders, including SZ and ASD [88, 89]. Hence, the observed functional enrichment profile highlights a likely pivotal role for the coordinated expression of the 3q29 interval genes and network partners that cluster in the turquoise module in establishing and maintaining the healthy functioning of the brain.

Functional characterization of the green module: Similar to the turquoise module, the pathway enrichment analysis of the genes constituting the green module revealed primary enrichment for shared function in several nervous-system specific biological processes. The overarching functional characteristics of this module are regulation of nervous system development, interactions between neuroactive ligands and receptors, synaptic vesicle cycle, intracellular trafficking systems (i.e., vesicle-mediated synaptic transport) and synapse assembly. The top Reactome-annotated biological pathways that were found to be enriched in the turquoise module include neuronal system (REAC:R-HSA-112316, adjusted  $P = 7.76\text{E-}07$ ), the role of GTSE1 in G2/M progression after G2 checkpoint (REAC:R-HSA-8852276, adjusted  $P = 6.14\text{E-}04$ ), potassium channels (REAC:R-HSA-1296071, adjusted  $P = 2.99\text{E-}03$ ), L1 cell adhesion molecule (L1CAM) interactions (REAC:R-HSA-373760, adjusted  $P = 8.69\text{E-}03$ ), transmission across chemical synapses (REAC:R-HSA-112315, adjusted  $P = 1.66\text{E-}02$ ), G protein-coupled receptor (GPCR) ligand binding (REAC:R-HSA-500792, adjusted  $P = 1.74\text{E-}02$ ), G alpha (q) signaling events (REAC:R-HSA-416476, adjusted  $P = 2.37\text{E-}02$ ), recycling pathway of L1 (REAC:R-HSA-437239, adjusted  $P = 3.93\text{E-}02$ ), coat protein complex I (COPI)-mediated anterograde transport (REAC:R-HSA-6807878, adjusted  $P = 4.27\text{E-}02$ ), and adenosine triphosphate-binding cassette

(ABC) transporter disorders (REAC:R-HSA-5619084, adjusted  $P = 4.36E-02$ ). The observed nervous system-specific functional enrichment findings suggest heightened disease-relevance for the 3q29 interval genes and intra-modular partners that coalesce in the green module. The neuropathology-associated functional characterization of the green module parallels that of the turquoise module, which shares the same meta-module as the green module. This functional overlap, which was identified agnostically to meta-module membership, presents further support for the utility of our approach in detecting biologically meaningful non-random network structures that organize gene expression in the healthy adult human cortex.

Functional characterization of the magenta module: The magenta module was found to be predominantly enriched for biological processes and pathways involved in post-translational protein modifications by small protein conjugation or removal, ubiquitin-dependent protein catabolism, intracellular protein transport and localization, and the ubiquitin-proteasome system. Hence, coordinated expression of the 3q29 interval genes and network partners that participate in the magenta module likely plays an important role in controlling the modification and spatiotemporal colocalization of substrates necessary for a variety of intracellular interactions. Moreover, pathway enrichment analysis revealed a link between magenta module genes and the initiation of MHC class I (MHC-I)-dependent immune responses, driven by a genomic locus that is increasingly implicated in the etiology of SZ [90]. MHC-I antigen presentation has been shown to strictly depend on peptide supply by the ubiquitin-proteasome system to initiate an effective adaptive immune response[91]; thus, the simultaneous enrichment of these interacting processes in a single module supports the biological relevance of the identified pattern of clustering. Specifically, the top Reactome-annotated biological pathways that were found to be significantly overrepresented in the magenta module include post-translational protein modification (REAC:R-HSA-597592, adjusted  $P = 1.22E-07$ ), gene expression (transcription) (REAC:R-HSA-74160, adjusted  $P = 7.98E-06$ ), MHC-I mediated antigen processing and presentation (REAC:R-HSA-983169, adjusted  $P = 1.69E-05$ ), antigen processing: ubiquitination and proteasome degradation (REAC:R-HSA-983168, adjusted  $P = 2.78E-05$ ), RNA polymerase II transcription (REAC:R-HSA-73857, adjusted  $P = 1.05E-04$ ), signaling by TGF-beta family members (REAC:R-HSA-9006936, adjusted  $P = 2.87E-03$ ), protein ubiquitination (REAC:R-HSA-8852135, adjusted  $P = 1.14E-02$ ), and sumoylation (REAC:R-HSA-2990846, adjusted  $P = 1.94E-02$ ). Overall, the functional profile of the magenta module encompasses many known regulators of brain function, including synapse formation and trans-synaptic signaling. Hence, the functional characteristics of the magenta module complement that of its meta-module partners, the green and turquoise modules.

Taken together, functional characterization of the 3q29 modules (Fig. 2d) point to novel mechanisms of shared or overlapping action for sets of 3q29 interval genes that cluster in the same network module and further coalesce in the same meta-module. Simultaneously, the variety of biological pathways that were found to be enriched in 3q29 modules suggests distributed involvement of this locus in not only nervous-system specific functions, such as regulation and organization of synaptic signaling and components, but also in core aspects of cell biology, including cellular metabolism, transcriptional regulation, protein modifications, and cell cycle regulation. Extended results of the pathway enrichment analysis are provided in Table S1.5.

Simultaneously, PPI network enrichment analysis revealed that all 3q29 modules show significant enrichment for PPIs that were systematically curated from the STRING protein interactome database (Fig. S8, Table S1.8), augmenting confidence in our RNA-Seq based network predictions with proteomic evidence (midnight blue, black, brown, and magenta modules:  $P$ -value  $< 1.00e-16$ ; dark turquoise module:  $P$ -value  $= 1.11e-16$ ; green module:  $P$ -value  $= 8.62e-08$ ; turquoise module:  $P$ -value  $= 4.30e-09$ ).

PPI Enrichment Statistics in 3q29 Modules   Source: STRING Protein interaction Database							
3q29 Module	Module Size	# of matching entries in STRING	Observed # of edges	Avg. node degree	Avg. local clustering coefficient	Expected # of edges	PPI enrichment p-value
Midnight blue	1,414	1,394	6,625	9.51	0.31	4,357	< 1.00E-16
Black	1,170	1,159	4,699	8.11	0.30	3,434	< 1.00E-16
Brown	1,972	1,928	10,902	11.30	0.29	7,930	< 1.00E-16
Magenta	1,437	1,415	7575	10.70	0.29	5,984	< 1.00E-16
Dark turquoise	496	479	911	3.80	0.33	683	1.11E-16
Green	2,000	1,928	8,887	9.22	0.29	8,403	8.62E-08
Turquoise	2,000	1,950	11,076	11.4	0.28	1,0481	4.30E-09

Detailed PPI enrichment statistics for each 3q29 module derived from the STRING Protein interaction Database are provided in the table above. A small PPI enrichment p-value indicates that the protein products of genes that were found to be highly co-expressed with 3q29 interval genes in our transcriptomic network analysis are not organized into modules at random and that the observed number of edges calculated for each 3q29 module based on PPI pairs curated from STRING is significant ( $P < 0.05$ ) (Fig. S8, Table S1.8). This enrichment analysis was complemented with other relevant PPI network statistics also listed in the table above, including average node degree, which reflects the number of intra-modular PPIs that the protein product of a gene from a given module has on average at the minimum required interaction score threshold of 0.4. The clustering coefficient is a measure of how connected the nodes in the inferred network are. The expected number of edges reflects how many edges are to be expected if the nodes were to be selected from the genome at random. More detail on active interaction sources and other parameters can be found in [47] and at <https://string-db.org/>."

Finally, we identified qualitative overlaps between the transcriptomic co-expression partners of 3q29 interval genes identified via WGCNA and known protein partners of 3q29 interval genes curated from the HuRI database (Fig. S7, Table S1.7). Of the 21 protein coding genes located in the 3q29 interval, only 14 (*CEP19*, *DLG1*, *FBXO45*, *MFI2*, *NCBP2*, *PAK2*, *PCYT1A*, *RNF168*, *SLC51A*, *TCTEX1D2*, *TFRC*, *UBXN7*) were found to have an entry on HuRI, 50% of which (*FBXO45*, *MFI2*, *NCBP2*, *RNF168*, *SLC51A*, *TCTEX1D2*, *TM4SF19*) had less than eight known proteome-wide interactors. A total of 193 distinct protein interactors were identified on HuRI for these 14 3q29 interval genes (after removing duplicates), 184 of which were identified as a node in our gene co-expression network. Of these 184, 137 (74%) were found to cluster in one of seven modules harboring 3q29 interval genes. 46% of the protein interactors identified in 3q29 modules share the same meta-module as their interacting 3q29 interval gene, 27% of which further show an overlap at the module level.

The full list of PPIs curated from HuRI and STRING, and brief statistics and visual illustrations of the resulting PPI networks can be found in Fig. S7, Fig. S8, Table S1.7 and Table S1.8.

### **Network modules found to harbor 3q29 interval genes are robust and strongly reproducible in an independent test dataset.**

To ensure the reproducibility of our network analysis results, we tested the preservation of various properties of graph structure that underly the modules identified in this study with respect to an

independent dataset obtained from the BrainSpan Project (Supp. Fig. S4). We calculated multiple density-based and connectivity-based preservation statistics for each module using a permutation test procedure and summarized the observed statistics by a composite Z-statistic,  $Z_{\text{summary.pres}}$ . All identified modules, except for the grey module (unassigned genes), were found to be successfully preserved in the test network ( $Z_{\text{summary.pres}} > 2$ ) (Fig. 1e). Specifically, 3/18 modules exhibited moderate evidence of preservation ( $2 < Z_{\text{summary.pres}} < 10$ ), and 15/18 modules, including all 3q29 modules, exhibited strong evidence of preservation ( $Z_{\text{summary.pres}} > 10$ ) (Fig. 1e). Moreover, the resulting composite preservation statistics of all 3q29 modules were substantially higher than that of a randomly drawn sample of 1,000 genes that represent the entire reference network as a single artificial module (labeled as the gold module,  $Z_{\text{summary.pres.gold}} = 6.78$ ).

In addition to preservation statistics, we calculated multiple module quality statistics that measure how well-defined or robust the boundaries of individual modules are in the reference network. By employing a resampling technique that applies module preservation statistics to repeated random splits of our reference data, we obtained a composite Z-statistic for each module ( $Z_{\text{summary.qual}}$ ) that standardizes and summarizes multiple cluster quality statistics. All 18 modules showed strong evidence for high cluster quality ( $Z_{\text{summary.qual}} > 10$ ), revealing robust module definitions (Supp. Fig. S5). Specifically, all 3q29 modules had a  $Z_{\text{summary.qual}}$  score  $\geq 20$  (Fig. 1e).

Finally, in line with the recommendations of the creators of the WGCNA package [39], we also evaluated the individual preservation and quality statistics underlying the composite  $Z_{\text{summary.pres}}$  and  $Z_{\text{summary.qual}}$  scores derived for each module. Individual module preservation statistics mostly converge on the finding that (1) nodes sharing the same module in the reference network remain highly connected in the test network and (2) connectivity patterns between nodes underlying the reference network remain similar in the test network (Supp. Fig. S5). Similarly, individual module quality statistics predominantly indicate strong evidence for high cluster quality in all identified modules across networks created from random splits of the original reference data (Supp. Fig. S5).

Overall, these findings support the strong reproducibility and robustness of our 3q29 modules, allowing high-confidence screening of the transcriptomic connectivity patterns formed by 3q29 interval genes in the healthy adult human PFC. See Table S1.4 for detailed results of the module preservation and quality analysis.

### **UBXN7 is a highly connected cortical hub-gene predicted to play a crucial role in the neuropsychiatric sequela of 3q29Del.**

To measure how strongly connected individual 3q29 interval genes are to their assigned network modules, we calculated the eigengene-based module connectivity measure (kME) of each 3q29 interval gene for its respective module (Fig. 2c). To reiterate, this measure quantifies how close a node is to its assigned module and can be applied to identify hub genes ( $kME > 0.8$ ,  $P < 0.05$ ), which are highly correlated with their module eigengene and exhibit high connectivity in their module. Intriguingly, our results revealed that *UBXN7*, an understudied and poorly annotated 3q29 interval gene, is a hub gene of its module ( $kME = 0.84$ ,  $P = 8.33E-30$ , midnight blue module size = 1,414 genes). Topological features of known disease-genes have been shown to disproportionately exhibit hub-gene characteristics compared to non-disease genes [53]. Supported by this literature, we predict that (1) *UBXN7* exerts central influence on a large network of co-expressed genes, and (2) loss of function mutations in *UBXN7* can cause major dysfunction in the biological pathways involving this gene. Consequently, we prioritize *UBXN7* as a major driver gene with likely disease relevance in 3q29Del. Intra-modular kME values of individual 3q29 genes, ranked from highest to lowest, are listed below:

<b><i>Eigengene-based module connectivity strength of 3q29 interval genes</i></b>			
<b>3q29 interval gene</b>	<b>Network module</b>	<b>Intra-modular kME</b>	<b>Associated p-value</b>
<i>UBXN7</i>	Midnight blue	0.84	8.33E-30
<i>SENP5</i>	Midnight blue	0.74	8.62E-20
<i>PAK2</i>	Dark turquoise	0.74	9.77E-20
<i>PIGX</i>	Turquoise	0.71	1.81E-17
<i>CEP19</i>	Turquoise	0.70	3.13E-17
<i>RNF168</i>	Magenta	0.70	5.06E-17
<i>NCBP2</i>	Brown	0.68	8.78E-16
<i>PIGZ</i>	Green	0.68	1.08E-15
<i>BDH1</i>	Green	0.64	8.32E-14
<i>FBXO45</i>	Turquoise	0.62	1.46E-12
<i>WDR53</i>	Midnight blue	0.59	2.04E-11
<i>DLG1</i>	Black	0.56	4.38E-10
<i>TCTEX1D2</i>	Turquoise	0.51	2.62E-08
<i>NRROS</i>	Turquoise	0.43	4.52E-06
<i>PCYT1A</i>	Green	0.40	2.00E-05
<i>TFRC</i>	Brown	0.39	3.56E-05
<i>ZDHHC19</i>	Brown	0.30	1.57E-03
<i>TM4SF19</i>	Brown	0.26	7.39E-03
<i>SLC51A</i>	Green	0.17	0.09
<i>SMCO1</i>	Green	0.11	0.25
<i>MFI2</i>	Green	0.09	0.35

Moreover, evaluation of the module membership strengths of 3q29 interval genes revealed that *SMCO1* (kME = 0.11,  $P = 0.25$ ), *SLC51A* (kME = 0.17,  $P = 0.09$ ) and *MFI2* (kME = 0.09,  $P = 0.35$ ) have non-significant kMEs for their assigned module, suggesting poor module connectivity. The mRNA expression summaries obtained from the Human Protein Atlas [92] (HPA) for 3q29 interval genes indicate nearly negligible or very low mRNA expression levels for *SMCO1* (consensus normalized expression value = 0.1), *SLC51A*, (consensus normalized expression value = 0.4), and *MFI2* (consensus normalized expression value = 2.8) in the human cerebral cortex. These data indicate the low abundance of these 3q29 interval genes in our tissue of interest, which likely relates to their peripheral network assignments in our analysis. Consequently, *SMCO1*, *SLC51A* and *MFI2* were excluded from downstream analysis to ensure accurate refinement of tight network connections formed by 3q29 interval genes.

A list of kME values and associated p-values for all network node-module pairs is provided in Table S1.3. A list of mRNA expression summaries obtained from the HPA for all 3q29 interval genes is provided in Table S1.6.

### **Nine 3q29 interval genes form transcriptomic subnetworks enriched for known SZ, ASD and IDD-risk genes.**

To systematically generate testable hypotheses regarding which 3q29 interval genes are causally linked to the major neuropsychiatric phenotypes associated with 3q29Del, we reduced 3q29 modules to strongly connected top neighbors of individual 3q29 genes and screened the resulting

top neighbors for a significant overlap with known SZ, ASD or IDD-risk genes. To reiterate, this approach leverages the extensively validated principle of guilt-by-association, which postulates that the disease-relevance of a particular gene is partially a property determined by its relationships in a biological network.

A top neighbor was defined as any node whose gene expression profile has a moderate-to-high pairwise correlation ( $\rho \geq 0.5$ ,  $P < 0.05$ ) with a 3q29 interval gene within the same module. Intriguingly, our results revealed that several 3q29 interval genes are among the top neighbors of one another within the same module. *FBXO45* ( $\rho = 0.5$ ,  $P = 5.43\text{E-}09$ ) and *PIGX* ( $\rho = 0.6$ ,  $P = 1.24\text{E-}10$ ) were identified as top-neighbors of *CEP19* in the turquoise module. Similarly, *SEN5* and *WDR53* were top-neighbors of each other ( $\rho = 0.5$ ,  $P = 1.05\text{E-}07$ ) in the midnight blue module. This finding further suggests that the correlated activity of subsets of 3q29 interval genes may converge upon the same or synchronized multicomponent biological processes in the adult PFC.

Moreover, *TM4SF19* (0 top neighbors) and *ZDHHC19* (3 top neighbors) were found to have no or  $< 5$  intra-modular partners in the brown module that met the correlation threshold to qualify as top neighbors. Similar to *SMCO1*, *SLC51A* and *MFI2*, the mRNA expression summaries obtained from the HPA[92] for *TM4SF19* (consensus normalized expression value = 0.5) and *ZDHHC19* (consensus normalized expression value = 0) indicate negligible or very low mRNA expression levels in the human cerebral cortex. These data independently indicate the low abundance of these 3q29 interval genes in our tissue of interest, which likely reflects the reason behind their lack of strongly connected top neighbors in our network analysis. Hence, *TM4SF19* and *ZDHHC19* were excluded from our downstream disease-association analysis, along with *SMCO1*, *SLC51A* and *MFI2*, which were deprioritized earlier due to poor module connectivity.

The intra-modular top neighbors of the remaining 16 3q29 interval genes were interrogated for overlap with six curated lists of evidence-based IDD, ASD or SZ-risk genes, spanning a wide range of the allele frequency spectrum (Fig. 3a). Hypergeometric test results, corrected for multiple testing, revealed a significant overrepresentation of one or more of these established risk gene-sets among the top neighbors of nine 3q29 interval genes (adjusted  $P < 0.05$ ): *BDH1*, *CEP19*, *DLG1*, *FBXO45*, *PIGX*, *RNF168*, *SEN5*, *UBXN7* and *WDR53* (Fig. 3b). Details of the enrichment results with respect to module membership are provided below.

In the black module, top neighbors of *DLG1* (gene-set size = 294) were found to be enriched for known SZ-risk genes from the exonic *de novo* mutations gene-set (adjusted  $P = 1.11\text{E-}05$ ). This identified intersection comprises 18 unique top neighbors of *DLG1* with known SZ association. Note that *DLG1* was the only 3q29 interval gene that clustered in the black module.

In the midnight blue module, all three constituent 3q29 interval genes had top neighbors that loaded highly onto known IDD, ASD and/or SZ-risk genes. Particularly, top neighbors of *UBXN7* (gene-set size = 811) were enriched for known IDD (overlap size = 14, adjusted  $P = 3.05\text{E-}04$ ), ASD (overlap size = 15, adjusted  $P = 5.46\text{E-}05$ ), and SZ-risk genes from the CommonMind case-control gene-set (overlap size = 45, adjusted  $P = 4.20\text{E-}03$ ). In addition, top neighbors of *SEN5* (gene-set size = 713) were enriched for known IDD (overlap size = 11, adjusted  $P = 5.05\text{E-}03$ ) and ASD-risk genes (overlap size = 12, adjusted  $P = 1.28\text{E-}03$ ). Similarly, top neighbors of *WDR53* (gene-set size = 278) had a significant overlap with known IDD (overlap size = 6, adjusted  $P = 1.51\text{E-}02$ ) and ASD-risk genes (overlap size = 6, adjusted  $P = 1.51\text{E-}02$ ). The union of these identified intersections adds up to a total of 67 unique top neighbors with known IDD, ASD and/or SZ association in this module.

In the green module, only two out of the six constituent 3q29 interval genes were found to have top neighbors that showed a significant overrepresentation of SZ-risk genes. Specifically, top neighbors of *BDH1* (gene-set size = 1008) were enriched for known SZ-risk genes from the CommonMind case-control gene-set (overlap size = 66, adjusted  $P = 4.49\text{E-}06$ ). Similarly, top neighbors of *PIGZ* (gene-set size = 995) were also enriched for known SZ-risk genes from the CommonMind case-control gene-set (overlap size = 58, adjusted  $P = 7.03\text{E-}04$ ). The union of these identified intersections adds up to a total of 75 unique top neighbors with known SZ association in this module.

In the magenta module, top neighbors of *RNF168* (gene-set size = 556) had a significant overlap with known IDD (overlap size = 9, adjusted  $P = 1.07\text{E-}02$ ) and SZ-risk genes from the CommonMind case-control gene-set (overlap size = 65, adjusted  $P = 5.21\text{E-}17$ ). The union of these identified intersections adds up to a total of 73 unique top neighbors with known SZ and/or IDD association in this module. Note that *RNF168* was the only 3q29 interval gene that clustered in the magenta module.

In the turquoise module, only two out of the five constituent 3q29 interval genes had top neighbors that loaded highly onto known IDD, ASD and/or SZ-risk genes. Specifically, top neighbors of *CEP19* (gene-set size = 1161) were enriched for known IDD (overlap size = 16, adjusted  $P = 1.64\text{E-}03$ ), ASD (overlap size = 15, adjusted  $P = 1.64\text{E-}03$ ), and SZ-risk genes from the exonic *de novo* mutations gene-set (overlap size = 29, adjusted  $P = 3.32\text{E-}02$ ). In addition, top neighbors of *FBXO45* (gene-set size = 1101) were also enriched for known IDD-risk genes (overlap size = 14, adjusted  $P = 1.35\text{E-}02$ ). The union of these identified intersections adds up to a total of 51 unique top neighbors with known IDD and/or SZ association in this module.

Finally, there was no statistically significant evidence for overrepresentation of known disease genes of interest among the top neighbors of 3q29 interval genes that clustered in the brown or dark turquoise modules.

To evaluate the specificity of the identified disease enrichment patterns, we also tested the top neighbors of 16 3q29 interval genes for overlap with known PD, late-onset AD and IBD risk genes. These disease phenotypes have no known link to 3q29Del syndrome, thus, genetic risk loci associated with these conditions were considered negative controls in this network analysis. In addition, a large list of common variants associated with height were included in our analysis as a fourth negative control to rule out a potential bias that could be introduced to our analysis by differences in the sizes of curated gene-sets. Our results indicate no statistically significant evidence for overrepresentation of AD or IBD-risk genes among the top neighbors of interrogated 3q29 interval genes. Only the top neighbors of *SENP5* (gene-set size = 713) showed a significant overlap with height-associated genes (overlap size = 30, adjusted  $P = 2.36\text{E-}02$ ). Additionally, the top neighbors of *NRROS* (gene-set size = 68), which did not show an enrichment for known IDD, ASD, or SZ risk genes, exhibited a small but significant overlap with known PD-risk genes (overlap size = 3, adjusted  $P = 2.00\text{E-}02$ ) (Fig. 3b).

Overall, 2 out of 64 hypergeometric tests indicated a significant overlap between the top neighbors of interrogated 3q29 interval genes and negative control gene-sets. In contrast, 19 out of 96 hypergeometric tests revealed a significant overrepresentation of SZ, ASD, and/or IDD-risk gene-sets among the same top neighbors, amounting to a proportion that is an order of magnitude larger than that of the negative controls (Fig. 3b). The substantial margin observed between these two enrichment ratios supports the high specificity and validity of our network-derived inferences for uncovering biology relevant to 3q29Del.

In summary, we identified 5,715 top neighbors of 3q29 interval genes, when combined across seven 3q29 modules. These top neighbors are predicted to function as direct interacting partners of 3q29 interval genes, participating in the same or overlapping biological pathways within the modular organization of molecular systems subserving the healthy functioning of the adult human PFC. Intriguingly, several 3q29 interval genes themselves were identified as top neighbors of other 3q29 interval genes, further suggesting functional convergence of subsets of genes within the 3q29 locus. Finally, our results revealed that *BDH1*, *CEP19*, *DLG1*, *FBXO45*, *PIGZ*, *RNF168*, *SENP5*, *UBXN7* and *WDR53* form strong co-expression-based ties with network partners that show a significant overlap with known SZ, ASD and/or IDD-risk genes curated from evidence-based literature (Fig. 3b). By leveraging the guilt by association principle, we prioritize these nine 3q29 interval genes, along with their 284 SZ, ASD, and/or IDD-related top neighbors from significant overlap tests as primary drivers of the major neuropsychiatric consequences of 3q29Del (Fig. 3b, Fig. 4a).

See Tables S2.1-3 for full lists of curated gene-sets, top neighbors, prioritized driver genes and detailed results of the overlap analysis.

### **Disease-relevant driver genes prioritized by network analysis load onto key biological pathways linked to neuropsychiatric disorders.**

To formulate testable hypotheses about the key biological mechanisms linking the 3q29 locus to major neuropsychiatric phenotypes associated with 3q29Del syndrome, we interrogated whether the prioritized driver genes identified in our network analysis assemble into known biological pathways and processes that are annotated in major gene ontology databases. Functional enrichment analysis on the union of our 293 prioritized driver genes (Fig. 4a) revealed their significant overrepresentation in several key biological pathways and processes, some of which are specific to nervous system function, while others are core cellular processes that are non-specific to an organ system (Fig. 4b, 4c).

Specifically, our findings indicate enrichment of our prioritized driver genes in eight biological pathways annotated by the Reactome and KEGG databases. These are axon guidance (REAC:R-HSA-422475, adjusted  $P = 3.64E-03$ ), post-translational protein modifications (REAC:R-HSA-597592, adjusted  $P = 5.24E-03$ ), long-term potentiation (KEGG:04720, adjusted  $P = 7.29E-03$ ), diseases of signal transduction (REAC:R-HSA-5663202, adjusted  $P = 1.00E-02$ ), regulation of actin cytoskeleton (KEGG:04810, adjusted  $P = 1.17E-02$ ), deubiquitination (REAC:R-HSA-5688426, adjusted  $P = 2.42E-02$ ), chromatin organization (REAC:R-HSA-4839726, adjusted  $P = 3.32E-02$ ), and diseases associated with glycosylation precursor biosynthesis (REAC:R-HSA-5609975, adjusted  $P = 4.06E-02$ ). This analysis also revealed the enrichment of our prioritized driver genes for several fundamental biological processes annotated by the Gene Ontology Project (GO:BP), including chromosome organization (GO:0051276, adjusted  $P = 3.81E-09$ ), histone modification (GO:0016570, adjusted  $P = 3.31E-08$ ), cellular component morphogenesis (GO:0032989, adjusted  $P = 4.57E-06$ ), regulation of organelle organization (GO:0033043, adjusted  $P = 6.40E-06$ ), DNA metabolic process (GO:0006259, adjusted  $P = 3.83E-05$ ), regulation of telomere maintenance (GO:0032204, adjusted  $P = 6.62E-05$ ), neuron differentiation (GO:0030182, adjusted  $P = 1.88E-04$ ), protein modification by small protein conjugation or removal (GO:0070647, adjusted  $P = 2.34E-04$ ), neuron projection morphogenesis (GO:0048812, adjusted  $P = 4.46E-04$ ), neurogenesis (GO:0022008, adjusted  $P = 1.89E-03$ ), chemical synaptic transmission, postsynaptic (GO:0099565, adjusted  $P = 2.37E-03$ ), post-embryonic development (GO:0009791, adjusted  $P = 2.66E-03$ ), synapse organization (GO:0050808, adjusted  $P = 3.20E-03$ ), protein acetylation (GO:0006473, adjusted  $P = 5.48E-03$ ), cell surface receptor signaling

pathway involved in cell-cell signaling (GO:1905114, adjusted  $P = 6.41\text{E-}03$ ), and excitatory postsynaptic potential (GO:0060079, adjusted  $P = 8.97\text{E-}03$ ). We hypothesize that the disruption of one or more of these biological pathways and processes, some of which have been demonstrated to be altered in idiopathic SZ and ASD [88, 89], lie on the casual pathway to neuropsychopathology in 3q29Del syndrome.

The top 20 biological processes and pathways enriched among our prioritized driver genes is shown in Fig. 4b. For clear illustration of our findings, we organized all identified GO:BP terms into a network of related functional annotation categories in Fig. 4c. Detailed results of this functional enrichment analysis, including a full list of prioritized drivers overlapping each identified gene ontology term are provided in Table S2.5.

### **Network-derived targets predict differentially expressed genes in the mouse model of 3q29Del.**

We tested the enrichment of the network targets identified in this study for differential expression in Del16<sup>+/-Bdh1-Tfrc</sup> mice compared with wild-type (WT) littermates [76]. RNA-Seq analysis revealed 290 protein-coding DEGs with known human homologs ( $P < 0.05$ ), 17 of which were identified as 3q29 interval genes (*Bdh1*, *Cep19*, *Dlg1*, *Fbxo45*, *Mfi2*, *Ncbp2*, *Nrros*, *Pak2*, *Pcyt1a*, *Pigx*, *Pigz*, *Rnf168*, *Senp5*, *Tctex1d2*, *Tfrc*, *Ubxn7*, *Wdr53*) (Fig. 5, Table S2.5). All 290 DEGs were tested for enrichment of network-derived targets identified via WGCNA at three scales of network interconnectedness: i) broad 3q29 network (11,924 genes), ii) top-neighbor-based 3q29 subnetwork (5,087 genes), and iii) prioritized drivers (280 genes). All compared gene-sets were filtered for mouse-human homology. Hypergeometric tests revealed significant enrichment of the interrogated DEGs for network-derived ties at all three levels of this analysis ( $P < 0.05$ ; Fig. 5b).

Specifically, 212 out of 290 DEGs were found to overlap with the broad 3q29 network ( $P = 2.42\text{e-}07$ ), with a 1.50-fold over-enrichment compared to what would be expected by random chance. 74 out of 290 DEGs were found to overlap with the top-neighbor-based 3q29 subnetwork ( $P = 0.03$ ), with a 1.22-fold over-enrichment compared to what would be expected by random chance. Finally, 12 out of 290 DEGs were found to overlap with prioritized drivers predicted to be associated with the neurodevelopmental and psychiatric consequences of 3q29Del ( $P = 1.43\text{e-}04$ ), with a 3.61-fold over-enrichment compared to what would be expected by random chance (Fig. 5b). The list of DEGs, including the subsets intersecting WGCNA-derived targets, and the list of genes corresponding to the three levels of network interconnectedness interrogated in this overlap analysis are provided in Table S2.5.

In conclusion, prediction of novel gene-function and gene-disease associations is an important goal in computational biology, particularly for un- or under-studied territories of the human genome, such as the recurrent 3q29Del locus. These genes have been neglected, in part, due to attention bias in biomedical research that disproportionately concentrates on isolated interrogation of well-studied genes [93]. The network-based guilt-by-association approach used in this study is a promising strategy to rectify this skew and to advance our understanding of the full complement of the human genome and the full scope of genetic risk for severe mental illnesses in a systems biology framework.

### **Availability of data & materials**

We provide two multi-sheet xlsx files (Supp. Tables S1 and S2), containing further detailed results of our analysis. Corresponding R code is available upon request.

The GTEx data (release version 6) used for the analyses described in this manuscript were downloaded from the GTEx portal on 01/08/2019 (<http://www.gtexportal.org/home/datasets/>, file name: "GTEx\_Analysis\_v6\_RNA-seq\_RNA-SeQCv1.1.8\_gene\_rpkm.gct.gz"); dbGaP accession number: phs000424.v6.p1. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

The BrainSpan data (Developmental Transcriptome Dataset) used to construct our test network were downloaded from the Allen Brain Atlas portal on 01/12/2019 (<https://www.brainspan.org/static/download/>, file name: "RNA-Seq Gencode v10 summarized to genes"); dbGaP accession number: phs000755.v2.p1.

## **Supplemental References**

1. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; **9**: 559.
2. de la Fuente A. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet* 2010; **26**(7): 326-333.
3. Mulle JG. The 3q29 deletion confers >40-fold increase in risk for schizophrenia. *Mol Psychiatry* 2015; **20**(9): 1028-1029.
4. Hafner H, Maurer K, Löffler W, Fatkenheuer B, an der Heiden W, Riecher-Rossler A *et al.* The epidemiology of early schizophrenia. Influence of age and gender on onset and early course. *Br J Psychiatry Suppl* 1994; (23): 29-38.
5. Howard R, Rabins PV, Seeman MV, Jeste DV. Late-onset schizophrenia and very-late-onset schizophrenia-like psychosis: an international consensus. The International Late-Onset Schizophrenia Group. *Am J Psychiatry* 2000; **157**(2): 172-178.
6. Fuster JM. Frontal lobe and cognitive development. *J Neurocytol* 2002; **31**(3-5): 373-385.
7. Owen MJ, O'Donovan MC. Schizophrenia and the neurodevelopmental continuum:evidence from genomics. *World Psychiatry* 2017; **16**(3): 227-235.
8. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering. *J Stat Softw* 2012; **46**(11).
9. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**(6235): 648-660.
10. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S *et al.* Functional organization of the transcriptome in human brain. *Nat Neurosci* 2008; **11**(11): 1271-1282.
11. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012; **22**(9): 1760-1774.
12. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR *et al.* Strategies for aggregating gene expression data: the collapseRows R function. *BMC Bioinformatics* 2011; **12**: 322.
13. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform* 2018; **19**(4): 575-592.
14. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012; **28**(6): 882-883.

15. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007; **8**(1): 118-127.
16. Ferreira PG, Munoz-Aguirre M, Reverter F, Sa Godinho CP, Sousa A, Amadoz A *et al.* The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat Commun* 2018; **9**(1): 490.
17. Wilcox R. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press: San Diego, 1997.
18. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004; **5**(2): 101-113.
19. Fu J, Keurentjes JJ, Bouwmeester H, America T, Verstappen FW, Ward JL *et al.* System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat Genet* 2009; **41**(2): 166-167.
20. Ouma WZ, Pogacar K, Grotewold E. Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties. *PLoS Comput Biol* 2018; **14**(4): e1006098.
21. Lachowiec J, Queitsch C, Kliebenstein DJ. Molecular mechanisms governing differential robustness of development and environmental responses in plants. *Ann Bot* 2016; **117**(5): 795-809.
22. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005; **4**: Article17.
23. Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet* 2001; **29**(4): 482-486.
24. Miller JA, Horvath S, Geschwind DH. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc Natl Acad Sci U S A* 2010; **107**(28): 12698-12703.
25. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 2007; **8**: 22.
26. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. *Science* 2002; **297**(5586): 1551-1555.
27. Carlson MR, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* 2006; **7**: 40.
28. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc Natl Acad Sci U S A* 2006; **103**(46): 17402-17407.

29. Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A* 2006; **103**(47): 17973-17978.
30. Ye Y, Godzik A. Comparative analysis of protein domain organization. *Genome Res* 2004; **14**(3): 343-353.
31. Li A, Horvath S. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics* 2007; **23**(2): 222-231.
32. Voigt A, Almaas E. Assessment of weighted topological overlap (wTO) to improve fidelity of gene co-expression networks. *BMC Bioinformatics* 2019; **20**(1): 58.
33. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008; **24**(5): 719-720.
34. Forsyth JK, Nachun D, Gandal MJ, Geschwind DH, Anderson AE, Coppola G *et al.* Synaptic and Gene Regulatory Mechanisms in Schizophrenia, Autism, and 22q11.2 Copy Number Variant-Mediated Risk for Neuropsychiatric Disorders. *Biol Psychiatry* 2020; **87**(2): 150-163.
35. Uddin M, Pellecchia G, Thiruvahindrapuram B, D'Abate L, Merico D, Chan A *et al.* Indexing Effects of Copy Number Variation on Genes Involved in Developmental Delay. *Sci Rep* 2016; **6**: 28663.
36. Gerring ZF, Gamazon ER, Derks EM, Major Depressive Disorder Working Group of the Psychiatric Genomics C. A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. *PLoS Genet* 2019; **15**(7): e1008245.
37. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M *et al.* Spatio-temporal transcriptome of the human brain. *Nature* 2011; **478**(7370): 483-489.
38. BrainSpan Atlas of the Developing Human Brain. Technical white paper: transcriptome profiling by RNA sequencing and exon microarray (v.5). 2013.
39. Langfelder P, Luo R, Oldham MC, Horvath S. Is My Network Module Preserved and Reproducible? *Plos Computational Biology* 2011; **7**(1).
40. Luxburg Uv. Clustering Stability: An Overview. *Foundations and Trends in Machine Learning* 2010; **2**(3): 235-274.
41. Miklos GL, Rubin GM. The role of the genome project in determining gene function: insights from model organisms. *Cell* 1996; **86**(4): 521-529.
42. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. *Nat Rev Genet* 2008; **9**(7): 509-515.
43. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* 2019; **14**(2): 482-517.

44. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* 2007; **35**(Web Server issue): W193-200.
45. Supek F, Bosnjak M, Skunca N, Smuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 2011; **6**(7): e21800.
46. Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W *et al.* A reference map of the human binary protein interactome. *Nature* 2020; **580**(7803): 402-408.
47. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019; **47**(D1): D607-D613.
48. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 2007; **1**: 54.
49. Cooper TF, Morby AP, Gunn A, Schneider D. Effect of random and hub gene disruptions on environmental and mutational robustness in *Escherichia coli*. *BMC Genomics* 2006; **7**: 237.
50. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* 2008; **322**(5898): 104-110.
51. Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 2005; **21**(23): 4205-4208.
52. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 2006; **22**(18): 2291-2297.
53. Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 2006; **22**(22): 2800-2805.
54. Sanati N, Iancu OD, Wu G, Jacobs JE, McWeeney SK. Network-Based Predictors of Progression in Head and Neck Squamous Cell Carcinoma. *Front Genet* 2018; **9**: 183.
55. Laine VN, Verhagen I, Mateman AC, Pijl A, Williams TD, Gienapp P *et al.* Exploration of tissue-specific gene expression patterns underlying timing of breeding in contrasting temperature environments in a song bird. *BMC Genomics* 2019; **20**(1): 693.
56. Li W, Wang L, Wu Y, Yuan Z, Zhou J. Weighted gene coexpression network analysis to identify key modules and hub genes associated with atrial fibrillation. *Int J Mol Med* 2020; **45**(2): 401-416.
57. Ala U, Piro RM, Grassi E, Damasco C, Silengo L, Oti M *et al.* Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput Biol* 2008; **4**(3): e1000043.

58. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. *Nat Genet* 2004; **36**(10): 1090-1098.
59. Torkamani A, Dean B, Schork NJ, Thomas EA. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. *Genome Res* 2010; **20**(4): 403-412.
60. van Dam S, Cordeiro R, Craig T, van Dam J, Wood SH, de Magalhaes JP. GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics* 2012; **13**: 535.
61. McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN *et al.* Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet* 2004; **36**(2): 197-204.
62. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 2017; **542**(7642): 433-438.
63. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism* 2013; **4**(1): 36.
64. Banerjee-Basu S, Packer A. SFARI Gene: an evolving database for the autism research community. *Dis Model Mech* 2010; **3**(3-4): 133-135.
65. Meng Q, Wang K, Brunetti T, Xia Y, Jiao C, Dai R *et al.* The DGCR5 long noncoding RNA may regulate expression of several schizophrenia-related genes. *Sci Transl Med* 2018; **10**(472).
66. Psych EC, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ *et al.* The PsychENCODE project. *Nat Neurosci* 2015; **18**(12): 1707-1712.
67. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* 2016; **19**(11): 1442-1453.
68. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; **511**(7510): 421-427.
69. Li J, Cai T, Jiang Y, Chen H, He X, Chen C *et al.* Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol Psychiatry* 2016; **21**(2): 290-297.
70. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet* 2019; **51**(3): 414-430.
71. Chang D, Nalls MA, Hallgrimsdottir IB, Hunkapiller J, van der Brug M, Cai F *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet* 2017; **49**(10): 1511-1516.

72. Momozawa Y, Dmitrieva J, Theatre E, Deffontaine V, Rahmouni S, Charlotiaux B *et al.* IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat Commun* 2018; **9**(1): 2427.
73. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* 2018; **27**(20): 3641-3649.
74. Shen L. GeneOverlap: Test and visualize gene overlaps. R package version 1.20.0 edn2019.
75. Chan SY, Loscalzo J. The emerging paradigm of network medicine in the study of human disease. *Circ Res* 2012; **111**(3): 359-374.
76. Rutkowski TP, Purcell RH, Pollak RM, Grewenow SM, Gafford GM, Malone T *et al.* Behavioral changes and growth deficits in a CRISPR engineered mouse model of the schizophrenia-associated 3q29 deletion. *Mol Psychiatry* 2021; **26**(3): 772-783.
77. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015; **31**(2): 166-169.
78. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15**(12): 550.
79. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**(1): 139-140.
80. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* 2017; **12**(12): e0190152.
81. Lin M, Pedrosa E, Hrabovsky A, Chen J, Puliafito BR, Gilbert SR *et al.* Integrative transcriptome network analysis of iPSC-derived neurons from schizophrenia and schizoaffective disorder patients with 22q11.2 deletion. *BMC Syst Biol* 2016; **10**(1): 105.
82. Brennand KJ, Simone A, Jou J, Gelboin-Burkhart C, Tran N, Sangar S *et al.* Modelling schizophrenia using human induced pluripotent stem cells. *Nature* 2011; **473**(7346): 221-225.
83. Sayers EW, Beck J, Bolton EE, Bourex D, Brister JR, Canese K *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2021; **49**(D1): D10-D17.
84. Varelas X, Miller BW, Sopko R, Song S, Gregorieff A, Fellouse FA *et al.* The Hippo pathway regulates Wnt/beta-catenin signaling. *Dev Cell* 2010; **18**(4): 579-591.
85. Moller LLV, Klip A, Sylow L. Rho GTPases-Emerging Regulators of Glucose Homeostasis and Metabolic Health. *Cells* 2019; **8**(5).

86. Madak-Erdogan Z, Charn TH, Jiang Y, Liu ET, Katzenellenbogen JA, Katzenellenbogen BS. Integrative genomics of gene and metabolic regulation by estrogen receptors alpha and beta, and their coregulators. *Mol Syst Biol* 2013; **9**: 676.
87. Su LF, Knoblauch R, Garabedian MJ. Rho GTPases as modulators of the estrogen receptor transcriptional response. *J Biol Chem* 2001; **276**(5): 3231-3237.
88. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* 2017; **49**(1): 27-35.
89. Guan J, Cai JJ, Ji G, Sham PC. Commonality in dysregulated expression of gene sets in cortical brains of individuals with autism, schizophrenia, and bipolar disorder. *Transl Psychiatry* 2019; **9**(1): 152.
90. Mokhtari R, Lachman HM. The Major Histocompatibility Complex (MHC) in Schizophrenia: A Review. *J Clin Cell Immunol* 2016; **7**(6).
91. Warnatsch A, Bergann T, Kruger E. Oxidation matters: the ubiquitin proteasome system connects innate immune mechanisms with MHC class I antigen presentation. *Mol Immunol* 2013; **55**(2): 106-109.
92. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 2015; **347**(6220): 1260419.
93. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol* 2018; **16**(9): e2006643.